Project Report: AI4Privacy

Dropped Neurons:
Anzellotti Alberto, Marenco Turi Gualtiero, Erifeloluwa Jamgbadi

27th October 2024

1 Abstract

Awareness of the importance of protecting Personally Identifiable Information (PII) and privacy-related data is essential to understanding the choices underlying our project. As a team, we view privacy as both a critical and captivating topic, fully supporting and valuing AI4Privacy's commitment to addressing these concerns. The rising risk associated with vast amounts of PII shared with chatbots and other AI-driven systems underscores the urgency of this project and the need for a timely solution, one that may lie within the AI domain itself.

2 Objective

With the Piiranha-v1 model achieving an impressive 98.27% performance in detecting PII, there was limited opportunity for improvement in detection accuracy. To further explore the potential of this benchmark model and future iterations, we developed a web-based user interface. This interface enables users to perform masking tasks using the selected model and view the model's confusion matrix.

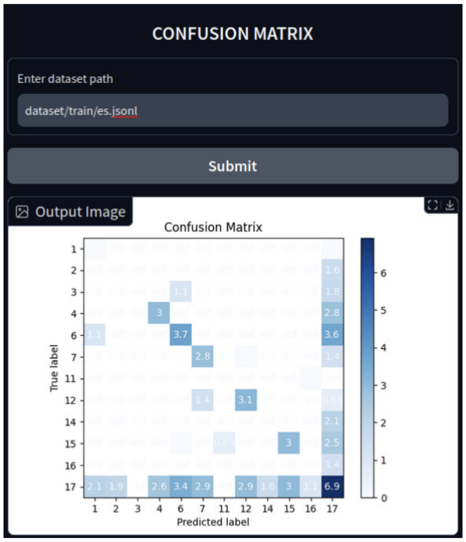
Initial analyses conducted with the Piiranha-v1 model highlighted several limitations, particularly in its suitability for multi-class classification tasks. Therefore, our objective was to develop a model that performs effectively across multiple PII classes. This new model aims to provide more specific and precise masking capabilities, alongside the ability to implement custom privacy policies tailored to diverse needs.

3 Approach

3.1 Finding Criticalities, Piiranha-v1 analysis

Our initial approach was to analyze Piiranha-v1 to understand the model's impressive performance. Since our goal was to create a benchmark for evaluating and testing future models as well, and we wanted this tool to be readily available to AI4privacy, we opted for a web-based user interface.This interface allows users to test the model's masking capabilities and view a confusion matrix of its predictions.

The confusion matrix was selected as a performance indicator due to its common use in classification tasks and ease of interpretability. For better scalability, we applied a natural logarithm transformation to the results.



CONFUSION MATRIX

Enter dataset path

dataset/train/es.jsonl

Submit

Output Image

Confusion Matrix

True label \ Predicted label	1	2	3	4	6	7	11	12	14	15	16	17
1		-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf	
2	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf	1.6
3	-inf	-inf	-inf	-inf	1.1	-inf	-inf	-inf	-inf	-inf	-inf	1.8
4	-inf	-inf	-inf	3	-inf	-inf	-inf	-inf	-inf	-inf	-inf	2.8
6	1.1	-inf	-inf	-inf	3.7	-inf	-inf	-inf	-inf	-inf	-inf	3.6
7	-inf	-inf	-inf	-inf	-inf	2.8	-inf		-inf	-inf	-inf	1.4
11	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf		-inf
12	-inf	-inf	-inf	-inf	-inf	1.4	-inf	3.1	-inf	-inf	-inf	0.69
14	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf	2.1
15	-inf	-inf	-inf	-inf		-inf	0.69	-inf	-inf	3	-inf	2.5
16	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf	1.4
17	2.1	1.9	-inf	2.6	3.4	2.9	-inf	2.9	1.6	3	1.1	6.9

Figure 1: Ln confusion matrix

Applying this approach to Piiranha-v1 revealed two main concerns:

- Masking Inefficiency: Piiranha's masking system produced suboptimal results in several tests. While classification performance was robust, the model's masking often failed to adequately obscure sensitive information in a proper form.
- Class Imbalance in Dataset: The confusion matrix indicated a dataset heavily skewed towards the "non-PII" class. Piiranha's impressive performance was largely due to accurate predictions in this dominant class, but it struggled to distinguish between different PII classes (e.g., passwords, email addresses, phone numbers), each representing distinct information types.

These insights prompted us to explore solutions aimed at enhancing Piiranha's ability to identify and mask various types of sensitive information. Distinguishing between different PII classes is critical not only for improved masking precision but also as a foundation for implementing custom privacy policies. For example, it might be desirable in a chatbot setting to retain an email address per user request, while ensuring that passwords are never stored—a privacy approach exemplified by systems like ChatGPT 4.0.

3.2 Foundation model and Meta-llama-3.1-8B-instruct

After evaluating several alternatives, we decided to use a foundation model—an advanced model trained on large datasets, which can be tailored ad hoc for specific applications. Specifically, we selected Meta's Llama-3.1, their most capable model to date, as described by Meta. Being open-source, Llama-3.1 has well-documented specifications readily accessible online.

Utilizing a large language model (LLM) offers numerous advantages, foremost among them the expectation that it understands the nuances of human language syntax and can assign different levels of significance to various PII classes.

4 Results

To evaluate the model's performance, we used a specialized scoring method, called "LLM as a judge" (RatherJudging LLM-as-a-Judge with MT-Bench and Chatbot Arena, Zheng et al.), as traditional metrics like confusion matrices are impractical for large language models (LLMs). This method is rather complex in practice, but the basic idea is to have a LLM evaluate another modelto compute a score for it's prediction capability. Our model's scored 0.71 against Piiranha's 0.70, considering that this result was achieved without fine-tuning and training only on 0.1 of the dataset due to time requirements, this is an appreciable result. This LLM can furthermore generalize better to new and zero shot token prediction scenarios. Moreover this LLM retain it's full ability to understand text so it can automatically parse constraints from policies and other contextual information. The model is applying appropriate masks and can go as far as to provide brief explanation on its action. Finally, we are confident in saying that we can implement a wide array of privacy policies which aim at treating different types of PII in not homogeneous way.

5 Challenges

As with any programming project, we encountered a range of issues. Beyond the typical syntax errors and minor mistakes, one of the biggest challenges was the limited size of our team. This constraint, combined with the novelty of the project for all of us, made the task both exciting and demanding; however, the limited manpower proved to be a significant obstacle.

In terms of technical difficulties, aside from being kicked out lighting.ai in the last hour of the project, one key issue was training the model for languages other than English. With more time, we could have employed a more comprehensive, albeit lengthier, training approach to better handle multilingual data. Nevertheless, we are confident that the English-trained model provides a reliable indication of the model's potential performance across other languages as well.